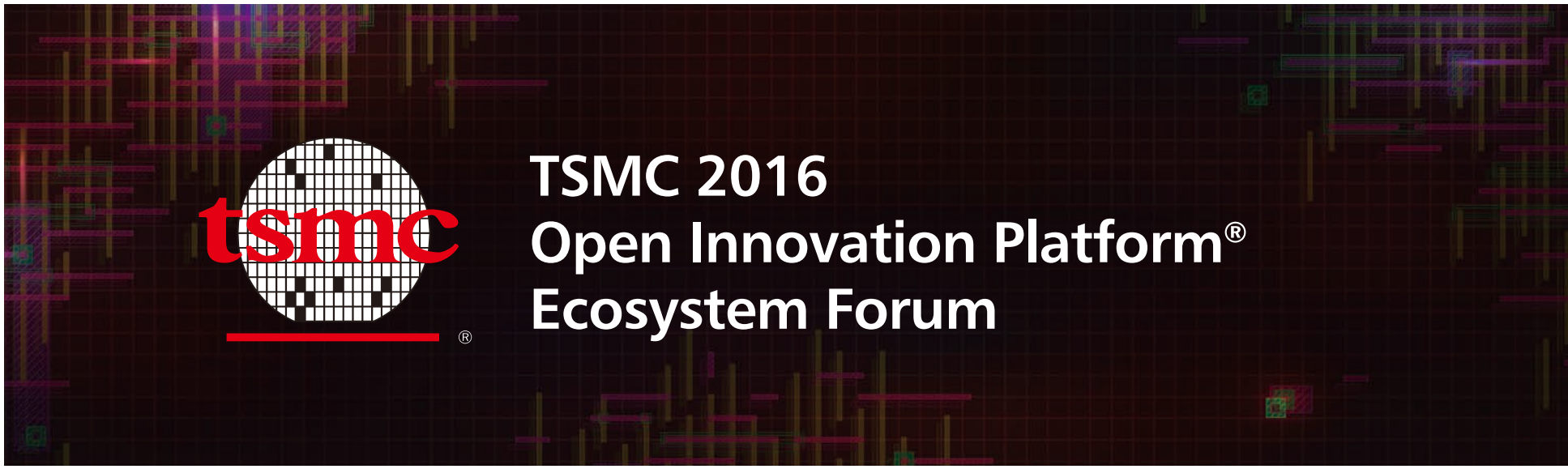
High Performance FinFET SRAM Design, Simulation, and Modeling

Synopsys



ABSTRACT

High speed circuit design is essential for high performance 64-bits server CPU. It is important for designers to obtain SPICE-accurate verification in order to reduce design margin and improve circuit performance. During the FinFET and multi-patterning era, circuit netlists are larger, parasitic are more complicate, and there are more process effect to be analyzed. The demand for fast yet accurate SPICE simulation remains extremely high.

In this presentation, we will present our experience of circuit simulation in high-speed SRAM design using Synopsys CustomSim. We will share the experience for accuracy & run-time trade-off in FinFET SRAM simulation. We will cover advanced modeling effects including self-heat and aging analysis, its usage experience in SPICE and how it impact design QoR.

Note: this presentation will be co-authored by R&D engineers from Oracle and TSMC.

Oracle Systems and Servers

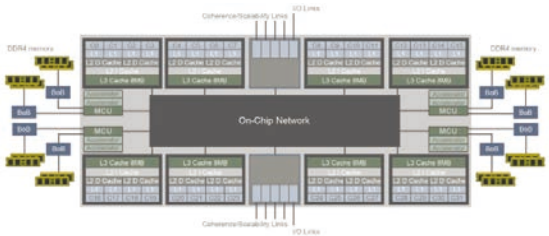
ORACLE SUPERCLUSTER

ORACLE SPARC SERVERS

Reference:
http://www.oracle.com/us/products/servers-storage

High Performance CPU

- SPARC M7 Processor
  - SQL in silicon
  - Data Analytic Accelerators (DAX)
- Design Highlights
  - 32 cores, 8 threads per core
  - 4.13 GHz, 20nm



Reference:
http://www.oracle.com/technetwork/server-storage/sun-sparc-enterprise/documentation/sparc-t7-m7-server-architecture-2702877.pdf

High Speed SRAM Design

- 64MB L3 cache sub-system supports application data integrity
- 256KB L2 data cache for each core pair
- SRAMs are optimized for power-speed trade-offs
- Multiple types of Vt devices are used to maximize leakage saving, while reducing dynamic power for high activity nodes
- Stacked PMOS are used for the last wordline driver stage to reduce leakage
- L3 data cache highly tuned for power and performance
- Dummy Transistors, and NMOS standby-on are also used

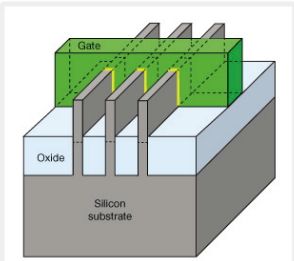
Reference:
- M7: Oracle's Next-Generation Sparc Processor, IEEE Mircro
- A 20nm 32-core 64MB L3 cache SPARC M7 processor, 2015 ISSCC

SRAM Simulation Using Fast-SPICE

- Fast-SPICE usage during past 15 years:
  - CustomSim (XA)
  - HSIM
  - StarSim-XT
- Microprocessor designed in:
  - Multiple technology nodes
  - Planar and FinFET device types
  - Different type of analysis including timing and power

Challenges for Simulating FinFET SRAMs

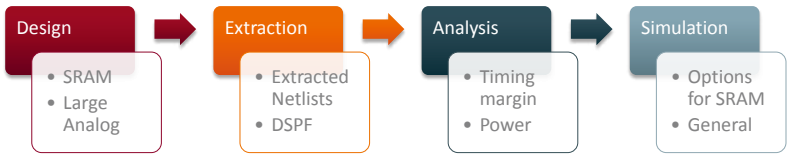
- Larger netlists
  - Polycide on Diffusion Edge devices (PODE)
  - More device parameters
  - More dummy transistors
  - More parasitic elements
- BSIM-CMG device model
- Self-heating device effects



Reference:
Synopsys' Modeling of 10 nanometer Parasitic Variation Effects Ratified by Open-Source Standards Board
https://www.synopsys.com/Company/Publications/DWTB/Pages/dwtb-finfet-process-soc-2015q1.aspx

Experience Using CustomSim

- CustomSim has been used for over 4 years (20nm node and FinFET)
- Custom WaveView used for viewing CustomSim results
- SRAM characterization options used extensively:
  - *set_sram_characterization*
  - *set_array_option*



SRAM Characterization Command

- Easy-to-use macro command that sets many minor commands
- SRAM characterization command results correlate well with SPICE simulation results
- *Sample usage:*
  - *Accurate timing measurements:*
    set_sram_characterization -application timing –accuracy 5
  - *Accurate current measurements:*
    set_sram_characterization -application power –accuracy 5

set_sram_characterization

- The Synopsys CustomSim set_sram_characterization command provides a single command that offers performance with high accuracy
  - Advanced RC reduction and optimization algorithm improves runtime while maintaining accuracy
  - Advanced memory circuit identification and partitioning algorithm to reduces partition size resulting in faster runtime
- It was developed to address the increased verification complexity of SRAM designs for the technology nodes of 16nm and below
  - Complex power management circuitry
  - Complex timer circuitry
  - Large amount of post-layout parasitic RC elements

Performance Benchmark

- Dramatic speedup as compared with set_sim_level on netlists that contain a large number of parasitics
  - Observed 3 to 30x speedup on extracted netlists.
  - Speedup dependent on number of MOSFETS and complexity of parasitic network.

SRAM block	Netlists size	Speed-up	Commands
A	14k+ MOS 494k+ Elements	25X	*set_sram_characterization*
B	244k+ MOS 5,992k+ Elements	20X+	*set_sram_characterization*
C	371k+ MOS 7,774k+ Elements	10X+	*set_sram_characterization*

Usage summary

- The runtime speedup of *set_sram_characterization* is significant in extraction netlists. There is minimal to no speedup on schematic-only pre-layout netlist.
- Also no run-time speedup if it run on a “cut” path, for example a reduced size netlist with a critical SRAM path only.
- For majority of our high-speed SRAM blocks, *set_sram_characterization* option is required. Otherwise it will take too long to run without this option (24+ hours).
- The options run fine for our SRAM blocks in both TSMC planar and FinFET process

Latest Design Challenges

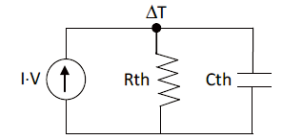
- Number of simulations
  - States (power save mode etc.)
  - Statistical variation
  - Corners
- Simulator capacity and accuracy
  - FastSPICE results should be compared against Golden SPICE when possible
  - DC solution is often the greatest source of discrepancy between simulators
- Self-Heat Effect (SHE)
  - Affected by Poly gate location
  - Introduced new instance parameters for poly gate location
    - Presence of parameters dramatically slowed down simulation (~6x) even when self-heat effect was not enabled (fixed in version vL-2016.03-1)

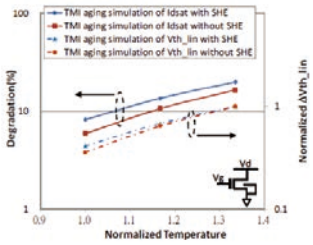
Challenge of SRAM simulation with Self-Heat

- Due to FinFET's 3D structure, self-heat is a device characteristic that require more sophistic modeling technique for accuracy and performance trade-off
- In order to accurately project delta-temperature increase of each FIN, fin location must be considered in SPICE netlists.
- New instance parameters are required in device definition.
- However these new SHE instance parameters breaks the typical algorithm of Fast-SPICE. Special handling is needed.

Self-Heat-Effect (SHE) modeling

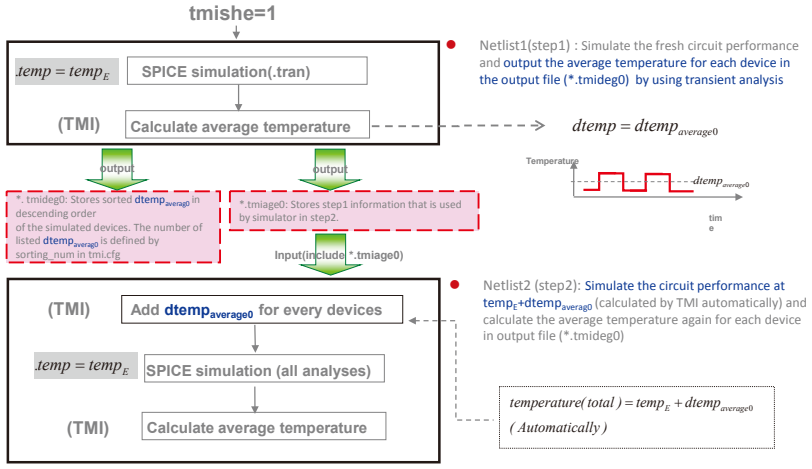
- Auxiliary R-C network
  - Used in most compact models including BSIM-CMG
  - Models thermal resistance (Rth) and capacitance (Cth)
  - Device branch current, nodal voltage, and device temperature are updated in each simulation iteration
  - Long simulation and potential convergence concern
- TSMC's TMI-based approach
  - Approximation approach to assume that device temperature raised by SHE is proportional to the average power (Pavg) dissipated in the device during a transient simulation period
  - Delta temperature is computed based on Pavg
  - No need to update device temperature for each iteration and time step
  - Little impact on simulation efficiency





Reference:
Unifying Self-heating and Aging Simulations with TMI2 (SISPAD 2014)

TMI Self-Heating Only Simulation Flow



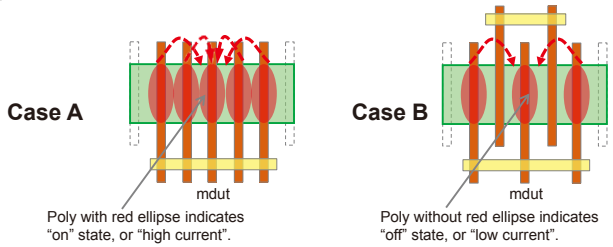
Netlist1(step1) : Simulate the fresh circuit performance and output the average temperature for each device in the output file (*.tmideg0) by using transient analysis

Netlist2 (step2): Simulate the circuit performance at $temp_E$+$dtemp_{average0}$ (calculated by TMI automatically) and calculate the average temperature again for each device in output file (*.tmideg0)

$temperature(total) = temp_E + dtemp_{average0}$
(Automatically)

New instance parameters for multi-finger effect

- Self-Heat Effect (SHE) model calculates increased temperature based on the environment of polys on the same OD
- **As is**: *dTemp* of **mdut** is as a function of multi-finger number on the same OD. Thus, *dTemp* of mdut in Case A is equal to that of Case B due to post-layout simulation instance od=5 for both cases.
- **To be**: *dTemp* of **mdut** is calculated from *dTemp* of polys on the same OD. New methodology is able to consider the contribution of low-power polys and therefore, *dTemp* of **mdut** in Case B is smaller than that of Case A.



Impact to Fast-SPICE algorithm and solution

- Device temperature increase due to SHE depends on environment of polys and OD.
- Fin location must be considered in SPICE netlists for accurate SHE effect modeling and simulation
- TSMC model introduced 3new instance parameters for Self-Heat-Effect (SHE) in 10nm, which impact SPICE capacity and performance

Parameters	Description
odind	index of each Oxide Diffusion (OD)
gpocrdx	x coordinate of each device center
gpocrdy	y coordinate of each device center

- The new solution maintains the simulator run-time and accuracy with the consideration of new SHE parameters

ORACLE

Summary

- CustomSim's SRAM characterization option has helped speedup SRAM simulations, we observed 3~30 times faster with acceptable precision to golden SPICE results.
- SRAM is a critical component of high-performance systems targeted at data center and cloud computing applications
- High-speed SRAM circuits require a fast, high-capacity and SPICE-accurate FastSPICE simulator
- FinFET netlists often contain in a larger number of parasitics than their equivalent planar versions
- FinFET technologies have introduced new effects modeling such as self-heating

ORACLE

Acknowledgement

- Oracle
  - Pouya Rezaeifakhr
  - Michael Yu
  - Jiajing Wang
- Synopsys
  - Zhaoping Chen
  - Joddy Wang
  - Weidong Liu

ORACLE

Hardware and Software
Engineered to Work Together

ORACLE

NOTE